PSTAT220A Homework 1

2025-10-13

Question 1 (25 pt)

Investigate the relationship between bodyweight, height and age in the data in `height_weight.csv`. Answer the following questions, supporting your answers with appropriate polished graphs.

a. Make a plots comparing height and weight, colored by age, and faceted by gender. Describe how weight varies as a function of height (e.g. is it approximiately linear, quadratic, etc?). Also comment on the variance of weight given age at different ages.

b. Find a transformation of weight that makes the relationship between height and weight more linear. Make the above plots replacing weight with the transformed weight.

c. At younger ages, the average heights of girls and boys are the same. At what age does the average height of boys diverge from the average height of girls? Support your answer with a clear and well constructed plot. Hint: you may want to use `geom_smooth` to more clearly visualize trends.

Question 2 (20 pt)

The following data are failure times in hours of 45 transmissions from caterpillar tractors belonging to a particular American company:

```
failure_times <- c(4381, 3953, 2603, 2320, 1161, 3286, 6914, 4007, 3168,
                   2376, 7498, 3923, 9460, 4525, 2168, 1288, 5085, 2217,
                   6922, 218, 1309, 1875, 1023, 1697, 1038, 3699, 6142,
                   4732, 3330, 4159, 2537, 3814, 2157, 7683, 5539, 4839,
                   6052, 2420, 5556, 309, 1295, 3266, 6679, 1711, 5931)
```

Use QQ-plots to examine the applicability of the following models for the probability distribution of failure time: normal, lognormal, exponential and Gamma (hint: check out the function `fitdistr` in the library MASS to fit these distributions. You may rescale the data if you have numerical problems). For the model that fits best (explain how you determine

which model fits best), plot the PDF and the kernel density estimate of the data on the same plot.

Question 3 (15 pt)

Generate 600 random samples from the normal distribution with mean 10 and standard deviation 5. Divide these 600 samples into 100 groups each with 6 samples. Compute the statistic $(\bar{X} - 10)/\sqrt{S^2/6}$ for each group. What kind of distribution do you expect this statistic to follow?

Using the 100 such statistics verify that the empirical distribution of these statistics actually follows the expected distribution.

Question 4 (40 pt)

This question considers the analysis of U.S. birth data. The data consists of the numbers of infants born in each month from 2016 through 2023 separated by state and race/ethnicity. The data also includes the number of women age 15-44 in each state and race/ethnicity (the `denom` variables). The prefix in the column variables indicates the race/ethnicity and the suffix indicates whether the variable tracks births or number of women.

```
birth_data <- read_csv("birth_data.csv", show_col_types = FALSE)
head(birth_data)
```

```
# A tibble: 6 x 14
  state  state_code  year month all_births white_births black_births hisp_births
  <chr>       <dbl> <dbl> <chr>      <dbl>        <dbl>        <dbl>       <dbl>
1 Alaba~          1  2016 Janu~       4805         2786         1462         381
2 Alaba~          1  2016 Febr~       4718         2846         1381         349
3 Alaba~          1  2016 March       4825         2839         1430         404
4 Alaba~          1  2016 April       4527         2730         1305         332
5 Alaba~          1  2016 May         4802         2908         1400         362
6 Alaba~          1  2016 June        5047         3077         1450         364
# i 6 more variables: all_denom <dbl>, white_denom <dbl>, black_denom <dbl>,
#   hisp_denom <dbl>, otherrace_births <dbl>, otherrace_denom <dbl>
```

a. Use `pivot_longer` to make the data tidy. Hint: you need to simultaneously pivot on `birth` columns and `denom` columns. To do so read the documentation for `pivot_longer` and use `.value` in the `names_to` argument (see e.g. the last example in the documentation). It might also help to use `names_sep`. After tidying, your data set should have 7 columns: `state`, `state_code`, `year`, `month`, `race`, `births` and `denom`. Print the first 10 rows and last 10 rows of your tidy data.

b. Create two new variable: `date` and `birth_rate`. To create `date`, use the `my` function from the `lubridate` package which takes a string consisting of the month followed by the year and returns the appropriate date object. `birth_rate` should have units of births per 1000 women 15-44 per year (note: don't forget to adjust for the fact that we have monthly data. As a reference, the national fertility rate is about 55 births per 1000 women aged 15-44 per year). Use the tidy data you just computed to plot the birth rate in California vs `date`. You should have lines with distinct colors for the birth rate in each race and overall. Which race/ethnicities tend to have the highest birth rates? Lowest?

c. Create a new variable using `mutate` which corresponds to the *relative* birth rate for each race category. The relative birth rate should be computed by grouping by race, taking the birth rate and dividing by the mean birth rate for that race over the full range of data and then ungrouping again. Plot the relative birth rates for all races on the same plot. What new observations do you make about the data when plotting the relative birth rate? Are any patterns clearer in this plot than they were in the previous plot?

d. Make a visualization of your choice that clearly highlights something about the data that was not evident in the previous plots. For example, you can explore seasonality in the trends or variation across race and/or states. Describe what you learned from your plot. The best visualizations will be shared in class (given your approval).